Ontologies étendues pour l'annotation sémantique*

Yue Ma, Laurent Audibert, Adeline Nazarenko

Laboratoire d'Informatique de l'université Paris-Nord (LIPN) - UMR 7030
Université Paris 13 - CNRS
99, avenue Jean-Baptiste Clément - F-93430 Villetaneuse, France
[prénom].[nom]@lipn.univ-paris13.fr

Résumé : Cet article tente de formaliser le processus consistant à annoter sémantiquement un texte au regard d'une ontologie. L'annotation sémantique met des fragments de texte en correspondance avec les éléments d'une ontologie, mais toute la difficulté consiste à identifier les fragments à annoter et les étiquettes à leur associer. Nous proposons d'étendre les ontologies par des règles d'annotation sémantique plutôt que par l'ajout de (méta-)propriétés lexicales. Cette solution permet de tirer le meilleur parti des outils de TAL qui produisent chacun à leur niveau des annotations linguistiques. Elle a également le mérite de distinguer clairement le processus d'analyse linguistique et l'interprétation ontologique.
Mots-clés : Ontologie, Annotation sémantique, Lexique, Terminologie, Plateforme d'annotation, Patron linguistiques

1 Introduction

L'annotation sémantique se définit comme le processus qui fixe l'interprétation d'un document en lui associant une sémantique formelle et explicite. Si cette interprétation s'exprime en termes ontologiques, nous parlons d'une interprétation ontologique. C'est le cas que nous considérons ici.

De nombreux systèmes annotent des textes au regard d'une ontologie qui peut être disponible localement ou via une URL (voir la revue proposée par (Dill *et al.*, 2003)). Il s'agit souvent d'annoter manuellement ou semi-automatiquement les texte (Handschuh, 2002; Kogut, 2001) mais les techniques d'apprentissage automatique permettent aussi d'automatiser l'annotation (Ciravegna, 2000).

Dans tous les cas, l'annotation repose sur des ontologies enrichies de connaissances linguistiques qui permettent de mettre en correspondance des éléments de l'ontologie avec des fragments de textes mais la nature de ces connaissances, leur encodage et leur mode d'utilisation varient beaucoup d'un système à l'autre. Nous proposons de

* Ce travail a été réalisé dans le cadre du programme Quaero, financé par OSEO, agence nationale de valorisation de la recherche.

les représenter sous la forme de règles d'annotation plutôt que d'adjoindre des propriétés lexicales aux concepts de l'ontologie. Alors que la plupart des systèmes d'annotation mettent l'accent sur l'étiquetage des entités nommées (Kiryakov *et al.*, 2004; Dill *et al.*, 2003), nous considérons le processus d'annotation sémantique dans toute sa complexité.

Une ontologie est une spécification formelle, explicite et consensuelle de la conceptualisation d'un domaine (Gruber, 1993). Une ontologie est constituée d'un ensemble de concepts organisés hiérarchiquement et structurés par des rôles liant ces concepts. Elle peut également comporter des axiomes et être peuplée. Dans ce dernier cas, elle comporte en outre des instances de concepts et des instances de rôles (nous parlons alors de relations entre instances)[1].

2 Annotation sémantique

L'annotation sémantique a pour objectif de formaliser l'interprétation qui peut être faite des textes sous la forme de méta-données attachées aux textes ou à certains de leurs segments. Cette interprétation s'exprime couramment en termes ontologiques quand il s'agit d'associer un type sémantique aux noms des entités mentionnées dans le texte (personnes, gènes, organisations, etc.) ou de les associer à un concept (Kiryakov *et al.*, 2004; Amardeilh *et al.*, 2005). Les entités nommées ne représentent cependant qu'une partie des éléments sémantiquement pertinents et de l'interprétation ontologique qui peut être faite des textes.

Pour illustrer notre propos, voici deux fragments de texte, chacun accompagné d'un exemple d'annotation sémantique respectivement représentés dans les figures 1 et 2 :

1. *Marie lit une pièce de théâtre de Molière.*
2. *The GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK (La protéine GerE inhibe la transcription in vitro du gène sigK qui encode sigmaK).*

2.1 Différents types d'annotation

Nous distinguons les types d'annotations selon la nature de l'élément ontologique auquel elles se rattachent.

Certains mots ou expressions renvoient à des *instances de concepts*. On les désigne traditionnellement sous le terme d'*entités nommées* car ils renvoient à des entités référentielles de manière autonome et conventionnelle (Ehrmann, 2008), comme les mots *Marie* et *SigmaK* ou l'expression *the GerE protein* dans les exemples ci-dessus. Si l'on considère des ontologies peuplées d'instances, le processus d'annotation consiste à créer des instances de concepts et à leur rattacher ces entités nommées. Si on interprète le texte au regard d'une ontologie classique (non peuplée), on néglige l'annotation des entités nommées ou on se contente de les associer à des concepts. Dans certains cas,

[1] Nous ne considérons pas ici de langage particulier pour la représentation des ontologies, mais dans la suite de l'article, les noms de concepts sont en majuscules (CONCEPT), les instances avec seulement une capitale à l'initiale (Instance), les rôles en minuscules (*a-pour-rôle*).

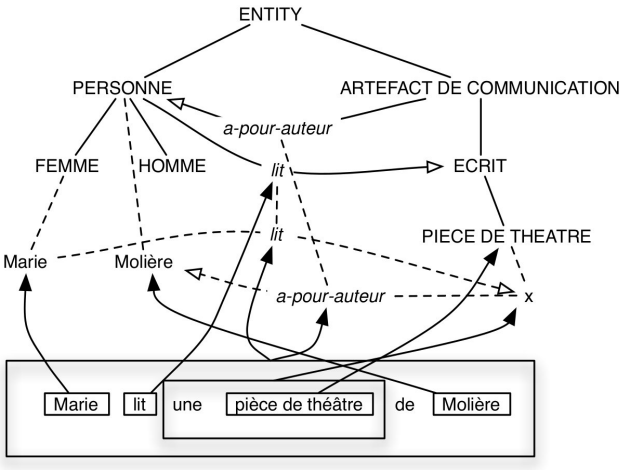

FIG. 1 – Example d'annotation de Marie lit une pièce de théâtre de Molière"

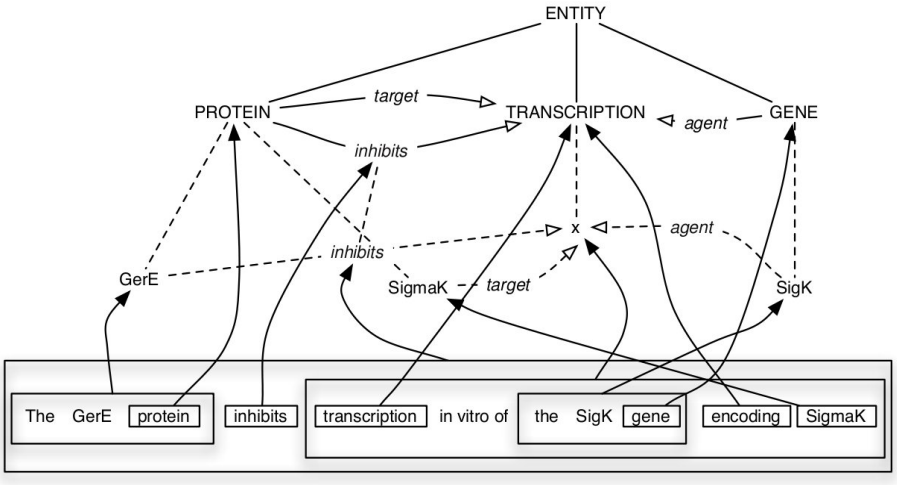

FIG. 2 – Example d'annotation de "The GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK
(La protéine GerE inhibe la transcription in vitro du gène sigK qui encode sigmaK)"

une référence est faite à une instance de concept sans que celle-ci soit nommée : dans l'exemple 1, la "pièce de théâtre de Molière" que "Marie lit" n'est pas mentionnée explicitement. Le processus d'annotation devrait être le même que lorsque l'entité est nommée mais en pratique on néglige ces "entités non nommées" parce que leur repérage est hors de portée des outils de traitement automatique des langues (TAL) courants.

Certains mots ou expressions dénotent des *concepts*. Ils constituent généralement le vocabulaire spécialisé du domaine considéré, *i.e.* la terminologie du domaine. Ces termes (par ex. *protein, transcription, pièce de théâtre*) sont souvent composés de plusieurs mots et ils sont importants à repérer parce qu'ils sont sémantiquement plus pertinents que les mots qui les composent. On a tendance à privilégier les termes nominaux mais des verbes ou groupes verbaux comme *encoding* peuvent aussi être associés à des concepts.

Certains mots ou expressions dénotent des *rôles conceptuels*. De même que les termes peuvent être rattachés à des concepts dans le processus d'annotation, ils peuvent être rattachés à des rôles si les notions sous-jacentes ont été modélisées sous la forme de rôle plutôt que comme concepts. C'est le cas de *lit* ou *inhibits* dans les exemples ci-dessus mais le fragment de texte annoté est souvent plus large, les rôles s'exprimant souvent par tournures de phrases qui ne se réduisent pas à un mot clef.

Certains fragments de texte renvoient à des *relations entre instances* : "une pièce de théâtre de Molière" ou "SigK gene encoding sigmaK" dans les exemples ci-dessus. C'est souvent un large fragment qui est annoté comme une relation entre instance. On s'y intéresse pour peupler les ontologies et on tend à les négliger lorsque l'interprétation se fait au regard d'une ontologie classique.

Certains fragments textuels, enfin, expriment des *axiomes ontologiques*. Par exemple, la phrase "Genes are biological entities" peut s'interpréter comme une relation de subsomption entre les concepts GENE et BIOLOGICAL ENTITY. Si on était capable d'analyser une phrase comme "les pièces de théâtre sont toujours écrites par quelqu'un", on pourrait de la même manière l'associer à un axiome exprimant une restriction de cardinalité du rôle *a-pour-auteur*.

Ces cinq types d'annotations ne sont généralement pas considérés tous ensemble. Selon les cas, on met l'accent sur la population des ontologies et donc sur les instances et leurs relations, sur l'information conceptuelle ou encore sur la découverte d'axiomes. Ils sont néanmoins intéressants à considérer ensemble pour appréhender l'annotation sémantique dans sa globalité.

2.2 Difficultés

Etablir les annotations précédentes pose une double difficulté.

La première difficulté concerne la segmentation du texte parce qu'il est souvent difficile d'identifier précisément les éléments textuels à annoter. C'est un problème connu en reconnaissance d'entités nommées : est-ce que le déterminant et le nom classifieur font partie de l'entité ou est-ce que seul le nom doit être annoté (*the GerE protein* vs. *GerE*) ? Il est souvent difficile de trancher. De manière plus générale, certaines connaissances ontologiques ne se traduisent pas par un simple mot ou expression. C'est souvent un fragment large ou une phrase complète qui véhicule l'information comme dans

le fragment "transcription of the SigK gene encoding sigmaK". Le fait de prendre ou non en compte ces annotations dépend largement de l'objectif visé.

L'ambiguïté inhérente à la langue soulève une seconde difficulté. Lorsque le fragment textuel est ambigu[2], il faut choisir l'élément ontologique avec lequel il doit être mis en correspondance. Le processus d'annotation suppose alors une étape de désambiguïsation qui repose généralement sur des indices figurant dans le contexte du fragment à annoter. Par exemple, la préposition *de* qui s'interprète comme dénotant le rôle *a-pour-auteur* dans "une pièce de théâtre de Molière", pourrait s'interpréter comme une localisation dans d'autres contextes. C'est le type des mots qui l'entourent et notamment le nom qu'elle introduit qui permet de désambiguïser la préposition.

Pour effectuer cette annotation sémantique en résolvant ces problèmes de segmentation et de désambiguïsation, il faut étendre l'ontologie avec des connaissances permettant de mettre en correspondance le texte à annoter et l'ontologie au regard de laquelle il est interprété.

2.3 Mise en oeuvre

Aujourd'hui, l'annotation linguistique des documents est généralement réalisée par des outils de TAL. On a souvent recours à des plates-formes qui permettent d'intégrer et d'exécuter, souvent séquentiellement, un certain nombre d'outils d'annotation existants qui ont été conçus dans des contextes et avec des objectifs différents. Parmi les plates-formes d'annotation linguistique, nous pouvons citer GATE (Cunningham, 2002), Atlas (Bird *et al.*, 2000) ou encore Ogmios (Hamon *et al.*, 2007). Une nouvelle tendance s'appuie sur l'environnement UIMA (Ferrucci & Lally, 2004).

Ces plates-formes prennent surtout en compte les problèmes d'interopérabilité verticale. Elles exploitent des mécanismes d'encapsulation des modules d'annotation, ce qui est une bonne solution si les annotations de ces outils sont fortement orthogonales.

Malheureusement, du point de vue de l'annotation sémantique, l'intersection entre les différents niveaux d'annotation n'est pas vide. Par exemple, dans le projet Alvis (Nazarenko *et al.*, 2006), l'annotation sémantique est en partie portée par le niveau de détection des entités nommées qui identifie ces entités et leur associe un premier type grossier, en partie par le niveau de détection des termes et en partie par un niveau spécifique d'annotation sémantique qui se limite à associer des catégories et des relations ontologiques aux éléments identifiés par les niveaux précédents. Ces trois niveaux sont donc fortement dépendants voire concurrents, ce qui pose d'importants problèmes de cohérence pour l'annotation sémantique.

Notre objectif ici est de résoudre ce problème d'interopérabilité horizontale non traité par les plates-formes d'annotation tout en tirant profit des solutions que ces dernières apportent au problème d'interopérabilité verticale. Il s'agit à la fois d'utiliser une ontologie pour rassembler en un lieu unique, cohérent et homogène ce qui est du ressort de la sémantique du document mais aussi d'étendre cette ontologie pour permettre son utilisation dans l'annotation sémantique, ce qui suppose de faire le lien entre le niveau conceptuel et le niveau linguistique qui sont par nature hétérogènes et non isomorphes.

[2]Le cas est fréquent : on sait que même dans les domaines spécialisés, les mots peuvent être ambigus.

Cette extension est généralement de nature lexicale (voir section 3) mais nous proposons une autre approche, à base de règles, qui préserve une articulation claire entre l'annotation linguistique (faite par les outils de TAL) et l'annotation sémantique qui est guidée par l'ontologie (section 4).

3 Les extensions lexicales

Il existe actuellement de multiples modèles pour représenter conjointement lexiques ou terminologies et ontologies. Dans cette optique, si OMV (*Ontology Metadata Vocabulary*) (Hartmann *et al.*, 2005) est une proposition de standardisation des méta-données descriptives d'une ontologie, LexOMV (Montiel-Ponsoda *et al.*, 2007) cherche à étendre le modèle OMV pour apporter des méta-données décrivant le niveau lexical des éléments (concepts ou propriétés) de l'ontologie. LexOMV n'est donc pas une extension lexicale mais permet de décrire, par des méta-données, l'extension lexicale d'une ontologie donnée, et donc de certains des formalismes que nous décrivons dans cette section. Dans un premier temps, nous nous intéressons aux formalismes, recommandés par le consortium W3C, que sont RDFS, SKOS et OWL. Ces trois formalismes sont finalement assez limités pour représenter correctement un niveau lexical riche. Nous décrivons donc ensuite quelques travaux qui cherchent à dépasser ces limitations avant d'avancer l'idée que l'annotation sémantique ne passe pas nécessairement par la représentation d'un niveau lexical riche au sein de l'ontologie.

RDFS est une recommandation du consortium W3C (W3C, 2009) qui permet de définir des étiquettes pour des classes via la propriété `rdfs:label`. Cette propriété peut être utilisée pour associer une information lexicale à une classe de l'ontologie. Mais le domaine de définition de `rdf:label` est le *Littéral*, ce qui limite l'expression d'informations lexicales complexes.

SKOS, actuellement en développement dans le cadre du consortium W3C, utilise les propriétés `skos:prefLabel` et `skos:altLabel` (sous-propriétés de `rdf:label`) et `xml:lang` pour associer des termes multilingues aux concepts. SKOS souffre des même limitations d'expression que RDFS pour rendre compte d'informations lexicales complexes. De plus, il faut noter que SKOS est conçu pour décrire une ressource conceptuelle et n'est pas adapté, contrairement à OWL, pour décrire la richesse structurelle des ontologies.

Toujours dans le cadre du consortium W3C, mais bien plus expressif que RDFS, OWL permet de formaliser des ontologies selon une syntaxe et une sémantique bien définies qui autorisent l'inférence. Même s'il se décline en trois sous-langages d'expressivité croissante (*OWL Light*, *OWL DL* et *OWL Full*), OWL n'offre pas plus de souplesse que RDFS pour la représentation du niveau lexical.

LingInfo (Buitelaar *et al.*, 2006), développé dans le cadre du projet SmartWeb, est un modèle de lexique basé sur une ontologie qui permet de représenter une terminologie multilingue. LingInfo associe les connaissances linguistiques aux classes (respectivement aux propriétés) de l'ontologie en définissant des meta-classes (resp. des meta-propriétés). La définition d'une meta-classe/propriété consiste en la donnée d'un terme, d'une langue et d'une décomposition morpho-syntaxique constituée d'un ensemble d'un ou plusieurs mots/syntagmes. Un syntagme est un mot/syntagme qui modélise

récursivement un syntagme nominal ou verbal complexe. Un mot est un mot/syntagme qui représente la structure morphologique de termes complexes comme *joueur de football*. A un mot peuvent être associées différentes propriétés comme le nombre ou le genre. Cette décomposition s'étend jusqu'à l'objet racine (*stem*) et permet donc d'associer aux éléments de l'ontologie une représentation lexicale très détaillée.

Szulman *et al.* (2008) proposent, avec le logiciel Terminae, une méthodologie de construction d'ontologies à partir de corpus. Le lien entre le corpus et les concepts est assuré par la construction d'une terminologie du domaine où chaque terme se traduit par une fiche terminologique. Terminae propose un export OWL et contourne les limitations de l'utilisation de la seule propriété `rdfs:label` par l'utilisation de la structure de propriété d'annotation (`owl:AnnotationProperty`) pour représenter les informations terminologiques associées à une classe.

Pour pallier la faiblesse d'OWL pour représenter le niveau lexical, (Reymonet *et al.*, 2007) et (Cimiano *et al.*, 2007) proposent une solution originale consistant à représenter le niveau lexical en utilisant toute l'expressivité d'OWL. Les termes sont ainsi réifiés sous forme de `owl:Class`. (Reymonet *et al.*, 2007) utilise un niveau d'abstraction supérieur pour distinguer les concepts des termes. Le lien entre concept et terme se fait par une propriété *denote* orientée du terme vers le concept. LexOnto (Cimiano *et al.*, 2007) va beaucoup plus loin et utilise une meta-ontologie pour mettre en relation le niveau lexical et le niveau ontologique. Ce formalisme permet de représenter des relations simples entre termes et concepts, mais également des structures linguistiques de type prédicat-argument comme les cadres de sous-catégorisation.

Les travaux décrits dans cette section présupposent la nécessité d'un niveau lexical ou terminologique pivot, entre la ressource ontologique et le texte. Ce niveau lexical pose deux problèmes importants, le premier est celui de sa représentation et le second celui de sa constitution. D'ailleurs, dans la plupart des travaux présentés ci-dessus, la constitution du niveau lexical est faite manuellement par un utilisateur assisté par un environnement logiciel et méthodologique adéquat. Pourtant, dans le cadre d'une plate-forme d'annotation linguistique, le niveau lexical existe déjà en grande partie. Le problème n'est pas tant son absence que la pluralité, l'hétérogénéité et l'incohérence des différents niveaux qui le supportent. Ainsi, plutôt que de reconstituer un nouveau niveau de représentation lexicale, nous pensons que notre extension d'ontologie doit plutôt s'appuyer sur les niveaux de représentation lexicale existants.

4 Une extension à base de règles d'annotation

Nous proposons de distinguer plus clairement les niveaux ontologique et lexical (ou linguistique) ainsi que le processus d'annotation sémantique des autres étapes d'annotation qui associent au texte des métadonnées linguistiques. Cela revient à établir une frontière claire et opératoire, même si elle est pour partie artificielle[3], entre l'interprétation du texte et son analyse linguistique.

En pratique, le processus d'annotation d'une plate-forme d'annotation linguistique se compose d'une série de modules d'annotation, chacun s'appuyant sur des ressources

[3]L'interprétation s'appuie sur l'analyse et celle-ci opère des choix qui sont en partie guidés par celle-là.

particulières (lexique, terminologies, etc.) pour ajouter une couche particulière d'annotations à partir des annotations des modules précédents, parfois en corrigeant ces dernières. L'ontologie étendue est exploitée comme une ressource exploitée par un composant d'annotation sémantique. L'extension de l'ontologie doit s'appuyer au maximum sur les niveaux inférieurs de la plate-forme d'annotation pour rester la plus simple possible et bénéficier du travail déjà réalisé par les composants inférieurs.

Cette extension doit être tout à la fois 1) compréhensible et modifiable par un être humain pour autoriser d'inévitables interventions manuelles sur une ontologie étendue, 2) interprétable par un ordinateur pour permettre l'annotation sémantique automatique dans le cadre d'une plate-forme d'annotation et 3) inférable par un ordinateur pour pouvoir être apprise (semi-)automatiquement à partir d'un corpus sémantiquement annoté selon l'ontologie à étendre. Il nous apparaît que la meilleure solution est l'utilisation de règles dont les prémisses sont constituées d'un ensemble de contraintes qu'un fragment de texte doit satisfaire pour être annoté par l'élément ontologique figurant dans la conclusion. L'application de la règle déclenche l'annotation de tous les fragments du texte qui satisfont les contraintes de la prémisse.

4.1 Définition de l'extension

Soit $O = \langle C, R, I?, RI?, A\rangle$ une ontologie composée d'un ensemble de concepts (C), de rôles (R), d'instances (I), de relations entre instances (RI) et d'axiomes (A)[4] et $\mathcal{R} = \langle \mathcal{R}_C, \mathcal{R}_R, \mathcal{R}_I, \mathcal{R}_{RI}, \mathcal{R}_A\rangle$, un ensemble de règles permettant d'annoter des fragments de textes en les reliant à des concepts ($\mathcal{R}_C$), des rôles ($\mathcal{R}_R$), des instances ($\mathcal{R}_I$), des relations entre instances ($\mathcal{R}_{RI}$) ou des axiomes ($\mathcal{R}_A$). Une règle est la donnée d'un couple (P, C) où P (*Prémisse*) décrit les conditions qu'un segment de texte doit vérifier pour être annoté et C (*Conclusion*) indique comment annoter le segment. Nous disons qu'une ontologie O^R est étendue ssi :

- pour chaque concept c de C il existe un couple de règles (ρ_c, ρ_{ci}) concluant sur c et telles que $\rho_c \in \mathcal{R}_C$ et $\rho_{ci} \in \mathcal{R}_I$;
- pour chaque rôle r de R il existe un couple de règles (ρ_r, ρ_{ri}) concluant sur r et telles que $\rho_i \in \mathcal{R}_I$ et $\rho_{ri} \in \mathcal{R}_{RI}$;
- pour chaque axiome a de A il existe une règle $\rho_a \in \mathcal{R}_A$ concluant sur a.

4.2 Types de règles

Comme indiqué ci-dessus, les règles sont différenciées selon le type de l'élément ontologique qui figure en conclusion de la règle mais il faut surtout distinguer deux types de règles :

- Les règles d'annotation ontologique à proprement parler ($\mathcal{R}_C$, $\mathcal{R}_R$, $\mathcal{R}_A$)concluent sur un concept, un rôle ou un axiome. Elles visent à identifier en corpus des fragments de texte qui dénotent des concepts, rôles ou axiomes et à les annoter en conséquence ;
- Les règles de peuplement ontologique ($\mathcal{R}_I$, $\mathcal{R}_{RI}$) concluent également sur des concepts ou des rôles mais elles visent à identifier des fragments de texte qui ren-

[4] Les éléments notés ? sont optionnels : ils n'apparaissent que dans les ontologies peuplées.

voient à des instances de ces concepts ou de ces rôles. Le fragment de texte est annoté comme une instance de concept ou de rôle et cette instance est ajoutée à l'ontologie sous le concept ou le rôle qui figure en conclusion de la règle.

Dans l'exemple de la figure 2, le segment *protein* est annoté par une règle du type ρ_c qui conclut sur le concept PROTEIN, tandis que le segment *The GerE protein* est annoté par une règle du type ρ_{ci} qui implique la création d'une instance GerE rattachée au concept PROTEIN.

Si l'objectif n'est pas de peupler l'ontologie, les règles de peuplement sont soit ignorées soit interprétées comme des règles conceptuelles. Dans ce cas, le fragment de texte qui renvoie à une instance est annoté comme le concept père de cette instance. Dans l'exemple de la figure 1, cela revient à annoter *Marie* non pas comme l'instance Marie mais comme le concept PERSONNE.

Concernant l'application des règles dédiées à l'identification d'occurrences d'instances, nous supposons que le nombre d'instances à créer correspond au nombre d'occurrences d'instances identifiées dans le texte. Le fait que deux occurrences dans le texte renvoient à un même individu est un problème de résolution de coréférence ou d'anaphore. Nous considérons que ce n'est pas du ressort de l'extension de l'ontologie de résoudre ce problème, qui doit trouver une solution à un autre niveau d'annotation de la plate-forme. Le fait de déléguer ce problème permet par ailleurs de centraliser les règles de peuplement au niveau du concept et de gagner en généricité. Dans le cas contraire, les règles devraient être réparties, spécialisées et attachées aux instances qui, contrairement aux concepts, ne sont pas des objets préexistants dans l'ontologie.

Ces extensions, associées aux concepts et aux rôles, ne sont pas des propriétés au sens où elles ne s'héritent pas. Le fait que le concept B hérite de A n'implique pas que les règles qui permettent d'identifier des fragments dénotant A puissent être utilisés pour identifier des fragments dénotant B.

4.3 Expression des règles

La prémisse d'une règle peut être représentée par un ensemble de patrons qui s'appliquent sur un corpus. S'il a été préalablement analysé par certains modules d'annotation (étiquetage morpho-syntaxique, reconnaissance d'entités nommées, étiquetage terminologique, par exemple) celui-ci porte déjà des annotations linguistiques. Un patron est une expression qui s'appuie sur ces différents niveaux d'annotations. L'application d'un patron sur un corpus est une opération qui retourne un ensemble de segments du corpus à annoter selon la conclusion de la règle.

A titre d'illustration, voici trois exemples distincts de patrons, écrits dans un pseudo langage pour en faciliter la compréhension, pour repérer dans le texte des occurrences du concept informatique *Système d'exploitation* :

1. `[string={système|d|'|exploitation}]`
2. `[lemme={système|de|exploitation}]`
3. `[terme={système d'exploitation}]`

string correspond à la forme brute du texte, *lemme* à la forme lemmatisée des mots et *terme* aux annotations de l'extracteur de termes. Ces trois patrons montrent l'intérêt de

l'utilisation des différents types d'annotations de la plate-forme. En effet, le premier patron n'est pas générique et ne peut pas reconnaître de simples variations comme *Système d'exploitation* ou *systèmes d'exploitation*. Le second est plus générique car insensible à la casse et au nombre. Le dernier est encore plus générique car, selon l'extracteur de termes utilisé, il peut reconnaître des chaînes comme *OS* pour lesquelles l'extracteur proposera la forme canonique *système d'exploitation*.

L'expression du patron ne peut pas toujours se réduire à la délimitation de la portion de corpus à annoter : il faut souvent exprimer des contraintes de désambiguïsation portant sur son contexte, comme pour les occurrences de *détention* qui sont, selon les cas, à annoter par un concept $\mathtt{D}_1$ correspondant au *fait d'être incarcéré ou enfermé*, ou par un concept $\mathtt{D}_2$ correspondant au *fait d'avoir en sa possession*. Pour englober ce cas de figure, un patron devrait plutôt s'écrire sous la forme d'un triplet $(LC?, T, RC?)$ où $LC?$ et $RC?$ sont des expressions facultatives pour contraindre le contexte gauche et droit, et où T (*Target*) est l'expression permettant d'identifier la portion de corpus qui doit supporter l'annotation.

Le langage utilisé pour l'écriture des patrons dépend de la façon dont le corpus et les annotations sont représentées. Dans le cas d'un document XML, un patron peut être la donnée d'un triplet d'expressions XPath ou d'une requête XQuery, l'expressivité de XQuery permettant de se passer de cette notion de triplet. Si les annotations et le corpus sont des objets Java persistants, les patrons peuvent, par exemple, se traduire par des requêtes JDOQL dans le cas d'une persistance gérée par JDO, ou des requêtes HQL dans le cas d'une persistance gérée par Hibernate.

4.4 Opérations sur une ontologie étendue

Étendre les ontologies suppose de redéfinir les opérations de mise à jour de l'ontologie, car toute opération effectuée sur l'ontologie a des répercussions sur son extension :

- La fusion de deux concepts est une opération simple. Soit c_1 et c_2 respectivement étendus par les couples de règles (ρ_{c1}, ρ_{i1}) et (ρ_{c2}, ρ_{i2}). On peut poser que le concept c résultant de la fusion de c_1 et c_2 a comme extension le couple de règles $(\rho_{c1} \vee \rho_{c2}, \rho_{i1} \vee \rho_{i2})$.
- La décomposition d'un concept impose la décomposition de son extension. Deux scénarios peuvent être envisagés. Le premier consiste à réaliser la décomposition de l'extension à la main par l'édition des règles associées à ce concept. Le second consiste à annoter un corpus représentatif avec le concept à décomposer en appliquant les règles qui lui sont attachées. L'utilisateur doit ensuite répartir les annotations sur les différents concepts qui remplacent le concept à décomposer. Le système peut ensuite reconstituer les extensions des différents concepts en inférant les patrons des règles à partir du corpus annoté.
- La suppression d'un concept peut se traduire par l'absorption de son extension par le ou les concepts qui le subsument, ce qui ramène au cas de la fusion. L'extension peut également être supprimée avec le concept : les fragments annotés cessent de l'être. Elle peut enfin être décomposée comme dans le cas précédent.
- L'ajout d'un nouveau concept implique de préciser son extension. Là encore deux scénarios sont envisageables : soit la proposition spontanée par l'utilisateur des

patrons associés, soit l'inférence de ces patrons à partir d'annotations manuelles des occurrences du concept créé dans un corpus représentatif.

5 Conclusion et perspectives

Cet article propose une solution au problème de l'annotation sémantique d'un texte au regard d'une ontologie. Nous nous interrogeons sur la manière dont des fragments de texte peuvent être mis en correspondance avec les éléments d'une ontologie. Nous avons proposé d'étendre les ontologies par des règles d'annotation sémantique plutôt que par l'ajout de (méta-)propriétés lexicales. Cette solution permet de tirer le meilleur parti des outils de TAL qui produisent chacun à leur niveau des annotations linguistiques. Elle a également le mérite de distinguer clairement le processus d'analyse et la tâche d'interprétation, qui seule met l'ontologie en jeu.

Étendre l'ontologie par des règles d'annotation qui s'expriment sous la forme de patrons présente, de notre point de vue, de nombreux intérêts. Tout d'abord, le pouvoir expressif des patrons est très grand. Il va de la simple représentation d'une liste de mots à des expressions complexes basées sur des annotations de haut niveaux (entités nommées, termes) et autorise l'expression de règles de flexion ou de désambiguïsation. Ensuite, les patrons sont compréhensibles et modifiables par une personne tout en étant interprétables, voire calculables, par un ordinateur, ce qui permet d'envisager leur acquisition automatique à partir d'un corpus annoté. Enfin, ils peuvent s'exprimer dans de nombreux formalismes largement connus voire standardisés comme les expressions régulières ou les chemins Xpath.

Ce travail appelle un double prolongement. Il s'agit tout d'abord de tester l'approche proposée en intégrant un module d'annotation sémantique dans une chaîne d'annotation existante. Nous projetons de le faire dans la plate-forme Ogmios et des expériences d'annotation sémantique doivent se faire dans le cadre du programme Quaero. La question de l'acquisition des règles est également une piste à explorer. Il est d'autant plus important de pouvoir apprendre (semi-)automatiquement les règles qu'elles varient d'une ontologie à l'autre mais aussi d'une plate-forme d'annotation à l'autre.

Références

AMARDEILH F., LAUBLET P. & MINEL J. L. (2005). Annotation documentaire et peuplement d'ontologie à partir d'extractions linguistiques. In *Actes des 16èmes journées francophones d'Ingénierie des Connaissances*, p. 25–36.

BIRD S., DAY D., GAROFOLO J. S., HENDERSON J., LAPRUN C. & LIBERMAN M. (2000). Atlas : A flexible and extensible architecture for linguistic annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, p. 1699–1706.

BUITELAAR P., SINTEK M. & KIESEL M. (2006). A multilingual/multimedia lexicon model for ontologies. In *ESWC*, p. 502–513.

CIMIANO P., HAASE P., HEROLD M., MANTEL M. & BUITELAAR P. (2007). Lexonto : A model for ontology lexicons for ontology-based nlp. In *Proceedings*

of OntoLex - From Text to Knowledge : The Lexicon/Ontology Interface (workshop at the International Semantic Web Conference).

CIRAVEGNA F. (2000). Learning to tag for information extraction from text. In *Proceedings of the ECAI-2000 Workshop on Machine Learning for Information Extraction*.

CUNNINGHAM H. (2002). Gate, a general architecture for text engineering. In S. NETHERLANDS, Ed., *Computers and the Humanities*, volume 36, p. 223–254.

DILL S., EIRON N., GIBSON D., GRUHL D., GUHA R., JHINGRAN A., KANUNGO T., MCCURLEY K. S., RAJAGOPALAN S., TOMKINS A., TOMLIN J. A. & ZIEN J. Y. (2003). A case for automated large scale semantic annotations. *Journal of Web Semantics*, **1**, 115–132.

EHRMANN M. (2008). *Les entités nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. Thèse de linguistique . Université de Paris VII.

FERRUCCI D. & LALLY A. (2004). UIMA : an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, **10**(3-4), 327–348.

GRUBER T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, **5**(2), 199–220.

HAMON T., DERIVIÈRE J. & NAZARENKO A. (2007). Ogmios : a scalable nlp platform for annotating large web document collections. In *Procedings of Corpus Linguistics 2007*, Birmingham, UK.

HANDSCHUH S. (2002). S-cream - semi-automatic creation of metadata. In *In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management, EKAW02*, p. 358–372 : Springer Verlag.

HARTMANN J., SURE Y., HAASE P., PALMA R. & DEL CARMEN SUÁREZ-FIGUEROA M. (2005). OMV – Ontology Metadata Vocabulary. In C. WELTY, Ed., *ISWC 2005 - In Ontology Patterns for the Semantic Web*.

KIRYAKOV A., POPOV B., TERZIEV I., MANOV D. & OGNYANOFF D. (2004). Semantic annotation, indexing, and retrieval. *J. Web Sem.*, **2**(1), 49–79.

KOGUT P. (2001). Aerodaml : Applying information extraction to generate daml annotations from web pages. In *First International Conference on Knowledge Capture (K-CAP 2001). Workshop on Knowledge Markup and Semantic Annotation*.

MONTIEL-PONSODA E., DE CEA G. A., SUAREZ-FIGUEROA M., PALMA R., PETERS W. & GOMEZ-PEREZ A. (2007). Lexomv : an omv extension to capture multilinguality. In *n Proceedings of the OntoLex07*, p. pp. 118–127.

NAZARENKO A., NÉDELLEC C., ALPHONSE E., AUBIN S., HAMON T. & MANINE A.-P. (2006). Semantic annotation in the alvis project. In *Proceeding of IIIA-2006 : International Workshop on Intelligent Information Access*, Helsinki, Finland.

REYMONET A., THOMAS J. & AUSSENAC-GILLES N. (2007). Modélisation de ressources termino-ontologiques en OWL. In F. TRICHET, Ed., *Journées Francophones d'Ingénierie des Connaissances (IC), Grenoble*, p. 169–180.

SZULMAN S., AUSSENAC-GILLES N. & DESPRES S. (2008). The Terminae Method and Platform for Ontology Engineering from Texts. In *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*, p. paru. IOS press.

W3C (2009). World wide web consortium. `http://www.w3.org/`.